Towards secure distributed healthcare research and delivery

Andrew Simpson
Software Engineering Programme
University of Oxford
February 22, 2007
Established in 1957, with a computing service being provided to the wider university and early research being focused on numerical analysis.

The Programming Research Group founded by Christopher Strachey in 1965.


Now a department with 45+ academics and a similar number of researchers.

As of January this year, the Laboratory is sharing space with the Life Sciences Interface Doctoral Training Centre and the Oxford e-Research Centre.
Research groups

- Computing Science
  - Theory and automated verification
  - Programming development tools
  - Applications and algorithms
- Numerical Analysis
- Software Engineering
The Software Engineering Programme

- The Software Engineering Programme is a joint venture between OUCL and the Department for Continuing Education
- The Programme began in the early 1980s as a set of ‘industrial courses’; an ‘integrated programme’ of six one-week courses was established in 1993
- This has evolved into a comprehensive programme of education in software engineering, covering 26 different courses: from discrete mathematics to web services

Open Day on March 3rd
The Software Engineering Programme is a part-time, *post-experience* programme of education: it teaches software engineering to people already in full-time employment.

All students are expected to have at least two years’ experience of large-scale software development, designing or programming as part of a development team.

The techniques taught in the Programme are thus taught, unavoidably, in the context of existing practice.
The design of the Programme

- Most industrial software engineering activity is project-based and, as a result, there are often times when students are simply unable to attend courses, or complete assignments; also, adult students may have other responsibilities that can interrupt their studies.
- To make the programme accessible, the taught component is delivered as a collection of residential, one-week modules, repeated once or twice a year, with a minimum of prerequisites or dependencies.
- During a teaching week, students are insulated, as far as possible, from the demands of the workplace.
Students on the Programme are drawn from a wide variety of backgrounds: SMEs, government agencies, large multinational corporations, etc.

Our teaching is informed by our involvement in the application of modern software engineering techniques and principles to large-scale software-based systems.

Hence our engagement with e-Research.

Design and security issues are of particular interest to us.
Projects

- e-DiaMoND
- climateprediction.net
- Integrative Biology
- NeuroGrid
- CancerGrid
- Integrative Biology Virtual Research Environment (IBVRE)
- IntBioSim
- National Cancer Research Institute (NCRI) prototype demonstrator for colorectal cancer care
- Generic Infrastructure for Medical Informatics (GIMI)
- ...
Collaboration

- All of these projects are large-scale, multidisciplinary, and collaborative
- Collaborators are drawn from other parts of the University (with the e-Research Centre being of significance in this respect), as well as other universities—Cambridge, Edinburgh, Loughborough UCL, KCL, Birmingham, Belfast, Auckland, Graz, Tulane, . . .
- and commercial partners: IBM, Microsoft, Siemens Molecular Imaging, T+ Medical, . . .
e-DiaMoND was a two-year 4.1 million pounds project funded through the EPSRC, the DTI, and IBM. The fundamental aim of the project was to deliver a prototype to demonstrate how emerging ‘grid’ technology could support the UK’s Breast Screening Programme. Applications were developed for computer-based training, epidemiological studies, and computer-aided detection. The project was led by Oxford and involved IBM, Mirada Solutions, UCL, KCL, and Edinburgh University—as well as the John Radcliffe Hospital, St George’s Hospital, Guy’s Hospital, and Edinburgh Breast Screening Unit.
The motivation

- One in four of all cancers in women are in the breast; men are also susceptible.
- It is estimated that one in eight women will develop breast cancer during their lives; it is also estimated that one in 28 women will die of the disease.
The project is developing an ‘integrative biology’ framework—facilitating modelling at various biological levels.

Initially this framework is being used for heart and cancer modelling.

Together, heart failures and cancer cause 60% of deaths in the UK.
Heart modelling

- Heart modelling requires intensive access to compute resources, significant data management facilities, visualisation capabilities, and collaborative working tools.
- Modelling typically involves solving coupled systems of PDEs (for the tissue level) and non-linear ODEs (for the cellular level) to model the electrical potential.
Cancer modelling

- The number and variety of mathematical models of solid tumour growth reflect the complexity of the phenomena involved.
- Most models focus on a single spatial scale of interest (sub-cellular, cellular, or organ)—even though there is compelling evidence that behaviour at these scales is closely linked.
- In addition, models typically describe generic—or idealised—tumours, rather than specific ones (such as breast or lung).
NeuroGrid

- Oxford-led three-year project, which commenced in March 2005
- The project aims to exploit web services, grid and e-Science technology to facilitate data sharing between neuro-imaging centres
- Current problems in neuro-imaging involve:
  - Poor data curation
  - Limited cross-fertilisation of ideas and techniques
  - Scanner differences
  - Movement artefact
NeuroGrid

- The development of middleware to enable secure data sharing
- The development of a work-flow engine and a toolkit
- Three ‘exemplars’
  - Dementia: carrying out a study to assess the value of web service-based applications in improving the consistency and analysis of imaging
  - Stroke: concerned with the development of an infrastructure for image management and observer interpretation in large-scale multi-centre studies
  - Psychosis: developing capabilities for remote analysis of existing data sets
The potential benefits of interdisciplinary research have been recognised for some time.
In recent years significant advances in technology have played an important role in the facilitation of interdisciplinary collaborations.
Researchers would like to:
- access information held at remote sites;
- aggregate data from disparate sources;
- control remote equipment;
- share compute resources;
- etc.
The main aims of the UK’s national e-Science Programme included:

- the building of a computational infrastructure to support large-scale research, and
- the identification of potential applications for such an infrastructure

Other initiatives have since been established in other countries, with Australia being a notable example

Within Oxford, the Oxford e-Research Centre is facilitating multi-disciplinary research—with the Software Engineering Programme being fully integrated into this activity
The Science and Innovation Investment Framework 2004–14

The framework, published in conjunction with the 2004 budget, states:

- “We need to enhance a culture of multidisciplinary research in the UK and provide the underpinning infrastructure to support it”
- “Over the decade many of the grand challenges in research will occupy the interfaces between the separate research disciplines developed in the 19th and 20th centuries”
Health grids

Health grids offer the potential for sharing of compute and data resources from different administrative domains to perform tasks that would otherwise be very difficult, if not impossible:

- compute grids offer the opportunity to provide unparalleled processing power to facilitate, for example, analysis of 3D images or real-time visualisation
- data grids offer the opportunity to share information between sites to allow distributed data analysis

But we could for some time about the term ‘grid’ . . .
The UK government has invested significant amounts (the initial estimate was approximately 6 billion pounds; the latest is approximately 12 billion pounds) in a National Programme for Information Technology (NPfIT) (since renamed Connecting for Health within England) in the NHS.

Similar schemes are being developed throughout Europe and in Australia, Canada and the United States to provide ‘cradle-to-grave’ views of patients via the linking of electronic information.
Electronic healthcare

- At the heart of CfH is the ‘spine’—the Care Record Service (CRS)—which is supposed to contain the electronic healthcare records of all patients in the country.
- The ‘electronic transfer of prescriptions’ service will allow prescriptions to be sent electronically to pharmacies.
- Postulated benefits include increased quality of care, increased efficiency, and increased patient autonomy.
- Research has been characterised as a ‘secondary use’—where, arguably, the real benefits will be realised.
Convergence

It seems almost inevitable that the aforementioned paths will converge in the near future, with real patient data stored in electronic patient records being used to support research.

Our projects are taking a long-term view in this respect.
Support for such use was recently voiced by the editors of the Journal of Medical Internet Research:

“Electronic records could facilitate new interfaces between care and research environments, leading to great improvements in the scope and efficiency of research. Benefits range from systematically generating hypotheses for research to undertaking entire studies based only on electronic record data . . . Clinicians and patients must have confidence in the consent, confidentiality and security arrangements for the uses of secondary data. Provided that such initiatives establish adequate information governance arrangements, within a clear ethical framework, innovative clinical research should flourish. Major benefits to patient care could ensue given sufficient development of the care-research interface via electronic records.”
The Secondary Uses Service

- SUS is a system to provide pseudonymised patient-based data
- Data provided for management and clinical purposes other than direct patient care
- ‘Secondary uses’ include health care planning, public health, clinical audit, research, and clinical governance
- There is also the provision of a range of tools to allow analysis
The Secondary Uses Service

- Data is supplied by care providers and submitted to a secure data transfer service (DTS)
- DTSs use XML for data interchange
- Data is stored in a central SUS repository
- Access is controlled by smartcard
- Users need to be registered to use the spine portal
- Access is enabled by local registration authority by adding a token with attributes to one’s smartcard
CfH: technical challenges [Becker, 2005]

- A lack of social controls that would otherwise prevent misuse of data
- The proposed spine is huge—containing records pertaining to 60 million+ patients
- Access rules are complex—and must reflect trade-offs between patient confidentiality, usability, and legislative constraints
- The requirements underpinning the system must be consistent with relevant legislation (the Data Protection Act, the Mental Health Act, the Human Fertilisation and Embryology Act, etc.)—which change often
The care record guarantee: some show-stoppers?

- “You can choose not to have information in your electronic care records shared”
- “If you are suffering distress or harm as a result of information being held in your record, you can apply to have the information amended or deleted”
- “You will be able to ask for a list of everyone who has looked at records about you and when they did so”
Some wider issues

- Are there inconsistencies with respect to thoughts on data ownership, data access and right to privacy (think: id cards, nightclub entry systems, “Megan’s law”, immigration, income tax, benefits, etc.)?

- Issues pertaining to procurement of systems: is there a need for education for customers?

- Does the need for “big bang” announcements lead to a lack of reality—and a failure to manage the public’s expectations—in the kinds of systems being procured?
Security considerations

- Distributed applications running across large-scale, decentralised, heterogeneous networks give rise to new security problems (and, arguably, give new life to old ones).
- In the healthcare context, we also need to think about:
  - Data ownership
  - Legal responsibility
  - Ethical treatment of data
  - Confidentiality and security of patient records
  - Appropriate anonymisation and pseudonymisation of data
  - Consent arrangements (whether ‘opt-in’ or ‘opt-out’)
  - Guaranteeing trust between practitioners and patients
  - Establishing workable governance arrangements
Security requirements derived from legislation

- We concern ourselves primarily with UK legislation
- Other countries have their own concerns to address: within the United States, for example, the Health Insurance Portability and Accountability Act of 1996 (HIPAA) is of concern:
  - the privacy rule dictates what uses and disclosures are authorised or required and patients’ rights with respect to their data
  - the security rule dictates what implementation is obligatory for enforcement of this policy or what reasonable efforts should be undertaken
The Data Protection Act

- Personal data shall be processed fairly and lawfully (and in accordance with certain conditions)
- Personal data shall be obtained for one or more specified and lawful purposes, and shall not be further processed in any manner incompatible with that purpose or those purposes
- Personal data shall be adequate, relevant and not excessive in relation to the purpose or purposes for which they are processed
- Personal data shall be accurate and, where necessary, kept up to date
The Data Protection Act

- Personal data processed for any purpose or purposes shall not be kept for longer than is necessary for that purpose or those purposes
- Personal data shall be processed in accordance with the rights of data subjects under the Data Protection Act
- Appropriate technical and organisational measures shall be taken against unauthorised or unlawful processing of personal data and against accidental loss or destruction of, or damage to, personal data
- Personal data shall not be transferred to a country or territory outside the European Economic Area unless . . .
The principles of the Caldicott Guardian

- Justify the purpose(s): every proposed use or transfer of patient-identifiable information within or from an organisation should be clearly defined and scrutinised, with continuing uses regularly reviewed by an appropriate guardian.
- Don’t use patient-identifiable information unless it is absolutely necessary.
- Use the minimum necessary patient-identifiable information.
The principles of the Caldicott Guardian

- Access to patient-identifiable information should be on a strict need-to-know basis
- Everyone should be aware of their responsibilities
- Understand and comply with the law: every use of patient-identifiable information must be lawful
Sharing and protecting data

- It should be noted that the principles of the Caldicott Guardian assume that each trust retains the ownership of all data located at its sites and each trust determines who can access its data (and under what circumstances): this is a model that we have adhered to when developing infrastructures to facilitate research.
- A key goal of our work is to design and deploy systems that allow data to be shared—while ensuring that the data owner retains absolute control over who can access which data, when it can be accessed, and even where it can be accessed from.
The UK is in an exceptional position to support the development of novel technical solutions to support individualised patient-specific healthcare.

Medical (and other) researchers often have significant amounts of data collected over a period of years, which commercial and research organisations (whether they are SMEs or large multi-nationals) would like to access.

Data sharing and data access are key to both.
Partners

- University of Oxford (Computing Laboratory and Engineering Science)
- University College London
- Loughborough University
- t+ Medical
- IBM UK
- Siemens Molecular Imaging
- The National Cancer Research Institute
The main aim of GIMI is to develop a generic, dependable middleware layer capable of:

1. (in the short term) supporting data sharing across disparate sources to facilitate healthcare research, delivery, and training;

2. (in the medium term) facilitating data access via dynamic, fine-grained access control mechanisms

3. (in the longer-term) interfacing with technological solutions deployed within the NHS via what might be termed an ‘ethical firewall’
Long-term conditions

- Within the UK, there is a drive towards self-management as a means of improving the health of patients with long-term conditions

- Self-management needs to be supported by a comprehensive IT system for disease management, which integrates all of the relevant information

- Initial research will focus on the development of robust algorithms for alerting healthcare professionals when the patient’s data deviate from the expected pattern
Mammography auditing and training

The main aim of this application is to demonstrate a prototype training tool for screening mammography which could offer radiologists a unique educational experience based on:

- Intelligent selection of training activities, based on knowledge both of the user’s performance, knowledge of radiological and pedagogical expertise
- A rich menu of training activities made possible through access to a very large number of sparsely annotated images
- A large number of richly annotated images
- A dataset annotated with specific learning goals
Medical imaging in cancer care

The overall aim of this application area is to:

- Develop medical image algorithms for application to breast and colorectal cancer
- Provide a testbed and user feedback for prototypes as they are developed throughout the duration of the project
Outline
OUCL and the Software Engineering Programme
Software Engineering research at Oxford
The emerging context
GIMI
For discussion . . .

Motivation
Partners
Aim
Applications
Key technical aspects
Architecture
Use cases
Research agenda
Technologies
Automatic build and deployment

Andrew Simpson  
BCS Oxfordshire Branch
Key technical aspects

- Federation
- Security
- Interoperability
We are concerned with intra-application, rather than inter-application federation.

To come up with an all-encompassing generic inter-application federation solution is beyond the scope of this project.
We are interested in security from two angles:

- Ensuring that our distributed infrastructure provides secure messaging, appropriate delegation mechanisms, acceptable authorisation and authentication mechanisms, etc.

- Providing fine-grained, dynamic access control policies that are user-defined so that data is shared on the user’s own terms
Interoperability

There are (at least) two aspects to consider here:

- Ensuring that our software is built using open source (or at least freely available) platforms and products and utilises commonly accepted standards
- With respect to the health domain, applications being HL7-compliant
A health grid architecture
A health grid architecture

- Each node contains a data store, externally facing services, internally facing services, access control policies, and workstations
- It is the externally facing services that allow different sites to communicate with each other
- Each site retains control of its data and access control policies
A health grid architecture

- All access requests are governed by policies accessible to the internal service, with the policies governing who can access the data, when they can access it, where it can be accessed from, and what rights they have to delegate that access.

- A local user makes a request to its local externally facing service, which then directs that request to the local internal service, another external service (or set of external services), or both the local internal service and other external services.
Use cases

- Distributed queries of patient data
- Working at a remote site
- Delegation of access permissions
- External access
- Modification of data
- Transferring patient records
Research agenda

- Fine-grained and evolving access control
- Distributed auditing
- Sophisticated delegation
- Techniques for ‘just enough’ anonymisation and query modification
The current system utilises:

- Linux (Gentoo) / IBM AIX for the server operating system
- Java 2 Standard Edition 5.0
- Apache Tomcat 5.0 and Java Web Services Development Pack 2.0
- Apache 2.2 using OpenSSL
- Apache Derby and IBM DB2 databases
- The Ant build tool
- The Eclipse development platform
The technologies we have chosen to create the GIMI infrastructure were chosen to ensure that it would be as portable and interoperable as possible: we have attempted to choose solutions with excellent cross-platform support and which is at least freely available—if not open source.
Automatic build and deployment

- Automated testing provides for build tests, system tests, and deployment tests
- The web services are automatically deployed using container management interfaces—we can instantly deploy updated services to a group of machines
- A health-check service provides a web-based status page that lets us know the status of all servers
Outline

OUCL and the Software Engineering Programme
Software Engineering research at Oxford
The emerging context
GIMI
For discussion...

GIMI Server Management

The server becomes gmserv01.diamond.ac.uk, and its alias is gmservm01.
There are 3 servers detected:

GIMI Server List

<table>
<thead>
<tr>
<th>No.</th>
<th>User Name</th>
<th>Creation Time</th>
<th>Expiry Time</th>
<th>Calculated Age</th>
<th>Used Space Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0108846C31B16E</td>
<td>Thu Feb 22 14:15:50 GMT 2007</td>
<td>Fri Feb 23 14:15:50 GMT 2007 00:00:00</td>
<td>01:00:00:00</td>
<td>0 KB</td>
</tr>
<tr>
<td>2</td>
<td>0108846C31B16E</td>
<td>Thu Feb 22 14:15:52 GMT 2007</td>
<td>Fri Feb 23 14:15:52 GMT 2007 00:00:00</td>
<td>01:00:00:00</td>
<td>0 KB</td>
</tr>
<tr>
<td>3</td>
<td>0108846C31B16E</td>
<td>Thu Feb 22 14:15:55 GMT 2007</td>
<td>Fri Feb 23 14:15:55 GMT 2007 00:00:00</td>
<td>01:00:00:00</td>
<td>0 KB</td>
</tr>
<tr>
<td>4</td>
<td>0108846C31B16E</td>
<td>Thu Feb 22 14:15:57 GMT 2007</td>
<td>Fri Feb 23 14:15:57 GMT 2007 00:00:00</td>
<td>01:00:00:00</td>
<td>0 KB</td>
</tr>
<tr>
<td>5</td>
<td>0108846C31B16E</td>
<td>Thu Feb 22 14:15:59 GMT 2007</td>
<td>Fri Feb 23 14:15:59 GMT 2007 00:00:00</td>
<td>01:00:00:00</td>
<td>0 KB</td>
</tr>
<tr>
<td>6</td>
<td>0108846C31B16E</td>
<td>Thu Feb 22 14:16:01 GMT 2007</td>
<td>Fri Feb 23 14:16:01 GMT 2007 00:00:00</td>
<td>01:00:00:00</td>
<td>0 KB</td>
</tr>
<tr>
<td>7</td>
<td>0108846C31B16E</td>
<td>Thu Feb 22 14:16:03 GMT 2007</td>
<td>Fri Feb 23 14:16:03 GMT 2007 00:00:00</td>
<td>01:00:00:00</td>
<td>0 KB</td>
</tr>
<tr>
<td>8</td>
<td>0108846C31B16E</td>
<td>Thu Feb 22 14:16:05 GMT 2007</td>
<td>Fri Feb 23 14:16:05 GMT 2007 00:00:00</td>
<td>01:00:00:00</td>
<td>0 KB</td>
</tr>
<tr>
<td>9</td>
<td>0108846C31B16E</td>
<td>Thu Feb 22 14:16:07 GMT 2007</td>
<td>Fri Feb 23 14:16:07 GMT 2007 00:00:00</td>
<td>01:00:00:00</td>
<td>0 KB</td>
</tr>
<tr>
<td>10</td>
<td>0108846C31B16E</td>
<td>Thu Feb 22 14:16:09 GMT 2007</td>
<td>Fri Feb 23 14:16:09 GMT 2007 00:00:00</td>
<td>01:00:00:00</td>
<td>0 KB</td>
</tr>
<tr>
<td>11</td>
<td>0108846C31B16E</td>
<td>Thu Feb 22 14:16:11 GMT 2007</td>
<td>Fri Feb 23 14:16:11 GMT 2007 00:00:00</td>
<td>01:00:00:00</td>
<td>0 KB</td>
</tr>
<tr>
<td>12</td>
<td>0108846C31B16E</td>
<td>Thu Feb 22 14:16:13 GMT 2007</td>
<td>Fri Feb 23 14:16:13 GMT 2007 00:00:00</td>
<td>01:00:00:00</td>
<td>0 KB</td>
</tr>
<tr>
<td>13</td>
<td>0108846C31B16E</td>
<td>Thu Feb 22 14:16:15 GMT 2007</td>
<td>Fri Feb 23 14:16:15 GMT 2007 00:00:00</td>
<td>01:00:00:00</td>
<td>0 KB</td>
</tr>
<tr>
<td>14</td>
<td>0108846C31B16E</td>
<td>Thu Feb 22 14:16:17 GMT 2007</td>
<td>Fri Feb 23 14:16:17 GMT 2007 00:00:00</td>
<td>01:00:00:00</td>
<td>0 KB</td>
</tr>
<tr>
<td>15</td>
<td>0108846C31B16E</td>
<td>Thu Feb 22 14:16:19 GMT 2007</td>
<td>Fri Feb 23 14:16:19 GMT 2007 00:00:00</td>
<td>01:00:00:00</td>
<td>0 KB</td>
</tr>
<tr>
<td>16</td>
<td>0108846C31B16E</td>
<td>Thu Feb 22 14:16:21 GMT 2007</td>
<td>Fri Feb 23 14:16:21 GMT 2007 00:00:00</td>
<td>01:00:00:00</td>
<td>0 KB</td>
</tr>
<tr>
<td>17</td>
<td>0108846C31B16E</td>
<td>Thu Feb 22 14:16:23 GMT 2007</td>
<td>Fri Feb 23 14:16:23 GMT 2007 00:00:00</td>
<td>01:00:00:00</td>
<td>0 KB</td>
</tr>
<tr>
<td>18</td>
<td>0108846C31B16E</td>
<td>Thu Feb 22 14:16:25 GMT 2007</td>
<td>Fri Feb 23 14:16:25 GMT 2007 00:00:00</td>
<td>01:00:00:00</td>
<td>0 KB</td>
</tr>
</tbody>
</table>

Andrew Simpson
BCS Oxfordshire Branch
Outline
OUCL and the Software Engineering Programme
Software Engineering research at Oxford
The emerging context
GIMI
For discussion . . .

GIMI Server Management
The starting server is gmovim12.chime.oucli.ac.uk, and its alias is gmovim13.

There are 2 servers detected:

**GIMI Server List**

1. Server Name: gmovim12.chime.oucli.ac.uk
   - Server Alias: gmovim12
   - The server has been up for approximately 30 minutes. (The number of minutes is updated every 10 minutes.)
   - File system used space: 2.1293 GB
   - Raw used space: 6.0 KB
   - File system free space: 142.068 GB
   - Raw free space: 142.068 GB

   [Folder information on this server]

2. Server Name: gmovim13.lboro.ac.uk
   - Server Alias: gmovim13
   - The server has been up for approximately 30 minutes. (The number of minutes is updated every 10 minutes.)
   - File system used space: 8.864 GB
   - Raw used space: 6.0 KB
   - File system free space: 288.809 GB
   - Raw free space: 288.809 GB

   [Folder information on this server]

You visited this page from host 143.1.189.37 on Thu Feb 22 14:04:00 GMT 2007.
We have developed—from scratch—a system based on Java and web services to support distributed healthcare research.

The reason for this was that—quite simply—the emerging ‘grid’ toolkit software wasn’t ready.
We use the Sun JWSDP, and, unfortunately, the WS-security libraries require extensions to the Java security libraries—which are shipped with all Sun-provided JVMs, but not IBM-provided JVMs.

It is arguable that the use of extensible standards to facilitate interoperability is doomed to failure as long as vendors are able to release product versions underpinned by their own closed-source extensions to these open standards.